

인공지능 학습용 데이터 품질검증 결과서

한국정보통신기술협회
소프트웨어시험인증연구소

주소: 경기도 성남시 분당구 분당로 47

결과서 관리번호:
DQ-2021-2-15-169



1. 데이터 명칭 : 고품질 과수 작물 통합 데이터
2. 데이터 구축 수행기관
 - 주관 : 지디에스컨설팅그룹 기업
3. 데이터 유형 : 이미지
4. 검증항목 : 통계적 다양성, 구문 정확성, 의미 정확성, 학습 유효성
5. 검증기간 : 2022. 1. 18. ~ 2022. 3. 4.

비 고 : 본 결과서의 결과는 데이터 구축 수행기관에서 제공한 데이터셋에 한하며 용도 이외의 사용을 금합니다.

확 인	작성자 성 명 : 남 사 욱	승인자
		직 책 : 기 술 책 임 자
		성 명 : 신 준 호

2022년 3월 4일

결과서번호 : DQ-2021-2-15-169

인공지능 학습용 데이터 품질검증 결과서:

고 품질 과 수 작 물 통 합 데 이 터



2022년 3월 4일

한국정보통신기술협회
Telecommunications Technology Association

본 문서는 한국정보통신기술협회(TTA) 소프트웨어시험인증연구소의 검증결과서로 한국정보통신기술협회(TTA) 승인 없이 문서의 전부 또는 일부를 발췌하여 배포, 복제 및 이용할 수 없습니다.

문서 관리 이력

버전	일자	변경 내용
1.0	2022.03.04	결과서 발행



목 차

1. 개 요	1
2. 검증항목 및 방법	3
2.1 검증항목	3
2.2 검증방법	4
3. 검증환경	5
3.1 데이터 검증조건	5
3.2 유효성 검증환경	6
4. 검증결과	7
5. 기타사항	9

1. 개 요

본 결과서는 한국정보통신기술협회(TTA)가 인공지능 학습용 데이터 구축 결과물을 대상으로 품질검증을 수행한 결과를 기술한 문서이다.

[표1] 품질검증 대상 데이터 요약

구분	내용																																				
데이터명	고품질 과수작물 통합 데이터																																				
구축목적	과수작물(감귤/키위) 병충해 이미지 촬영 및 Annotation 가공을 통하여 감귤의 병충해 3종과 키위의 병충해 3종의 분류를 위한 AI 학습용 데이터셋을 구축																																				
라벨링 방법	<ul style="list-style-type: none">◦ 클래스 라벨링(채집 이미지)<ul style="list-style-type: none">- 채집하여 대상 객체와 구별되는 배경으로 촬영된 객체는 바운딩박스, 폴리곤 등의 어노테이션 구현하지 않고 라벨링 함.- 라벨은 과수별로 정상, 병충해명의 Class로 분류하여 부여함.◦ 폴리곤 어노테이션(현장 이미지)<ul style="list-style-type: none">- 과수작물의 잎, 과실을 촬영한 후 잎, 과실에 대해 폴리곤으로 작업- 폴리곤은 해당 객체를 완전하게 포함하도록 어노테이션- 폴리곤의 라벨은 과수별로 정상, 병충해명의 Class로 분류하여 부여함.																																				
데이터 종류/형식	<ul style="list-style-type: none">• 원천 데이터: 이미지(.JPG)• 라벨 데이터: .json (어노테이션 폴리곤이 있는 경우 포함)																																				
클래스 수량	<p>감귤: 병충해별(정상, 궤양병, 굴응애, 진딧물) 4종*</p> <p>키위: 병충해별(정상, 점무늬병, 총채벌레, 과실무름병) 4종*</p> <p><표. 병충해 클래스별 수량></p> <table><tr><th>클래스</th><th>과수명</th><th>병충해</th><th>수량</th><th>비고</th></tr><tr><td rowspan="4">이미지 (40,000장)</td><td rowspan="2">감귤 (20,000장)</td><td>정상</td><td>5,000</td><td>정상 이미지</td></tr><tr><td>병충해</td><td>15,000</td><td>궤양병, 굴응애, 진딧물</td></tr><tr><td rowspan="2">키위 (20,000장)</td><td>정상</td><td>5,000</td><td>정상 이미지</td></tr><tr><td>병충해</td><td>15,000</td><td>점무늬병, 총채벌레, 과실무름병</td></tr></table> <p>(※ 매년 병충해 발생 상황의 변동에 따라 부위별 구분 없이 병충해별 비율임.)</p> <p>* 병충해별 클래스 구분은 <첨부3. 클래스 체계 분류> 참조</p> <p>감귤 : 생육단계별</p> <p>키위 : 품종별</p> <p><표. 초분광 이미지 수량></p> <table><tr><th>구분</th><th>과수명</th><th>생육/품종</th><th>수량</th><th>비고</th></tr><tr><td rowspan="4">초분광 이미지 (100,000장)</td><td rowspan="2">감귤 (50,000장)</td><td>성장기</td><td rowspan="2">50,000</td><td rowspan="2">분포확인</td></tr><tr><td>성숙기</td></tr><tr><td rowspan="2">키위 (50,000장)</td><td>골드키위</td><td rowspan="2">50,000</td><td rowspan="2">분포확인</td></tr><tr><td>레드키위</td></tr></table>	클래스	과수명	병충해	수량	비고	이미지 (40,000장)	감귤 (20,000장)	정상	5,000	정상 이미지	병충해	15,000	궤양병, 굴응애, 진딧물	키위 (20,000장)	정상	5,000	정상 이미지	병충해	15,000	점무늬병, 총채벌레, 과실무름병	구분	과수명	생육/품종	수량	비고	초분광 이미지 (100,000장)	감귤 (50,000장)	성장기	50,000	분포확인	성숙기	키위 (50,000장)	골드키위	50,000	분포확인	레드키위
클래스	과수명	병충해	수량	비고																																	
이미지 (40,000장)	감귤 (20,000장)	정상	5,000	정상 이미지																																	
		병충해	15,000	궤양병, 굴응애, 진딧물																																	
	키위 (20,000장)	정상	5,000	정상 이미지																																	
		병충해	15,000	점무늬병, 총채벌레, 과실무름병																																	
구분	과수명	생육/품종	수량	비고																																	
초분광 이미지 (100,000장)	감귤 (50,000장)	성장기	50,000	분포확인																																	
		성숙기																																			
	키위 (50,000장)	골드키위	50,000	분포확인																																	
		레드키위																																			
데이터 예시	<ul style="list-style-type: none">• 원천 데이터																																				



• 라벨 데이터

```
{
  "Info": {
    "IMAGE_FILE_NM": "HF01_00FT_000001",
    "IMAGE_FILE_TY": "JPEG",
    "STRE_COURS": "1.Training/라벨링데이터/01.감귤,
                  1.Training/원천데이터/01.감귤",
    "RSOLTN": "(1920,1080)",
    "IMAGE_POTOGRF_DT": "2021:08:05 16:21:12",
    "CMRA_INFO": "samsung",
    "F_STOP": "2.0",
    "FILM_SPD": "45",
    "FLSLT_USE_AT": "N",
    "LCINFO": "F02",
    "IMAGE_OBTAIN_PLACE_TY": "노지",
    "GRWH_STEP_CODE": "6",
    "DBYHS": "00",
    "OCPRD": "08-05",
    "SPECIES_NM": "온주밀감"
  },
  "Annotations": {
    "ANTN_ID": null,
    "ANTN_TY": null,
    "OBJECT_CLASS_CODE": "감귤_정상",
    "BNDBX": null
  },
  "Environment": {
    "ENVIRON_INFO_FILE_NM":
    "HF01_F01_20210801-20211130.csv",
    "MSRG_DATETM": "2021-08-05 16:21:12",
    "SOLRAD_QY": null,
    "AFR": null,
    "TP": "OOO",
    "HD": null,
    "SOIL_MITR": null
  }
}
```

2. 검증항목 및 방법

2.1 검증항목

인공지능 학습용 데이터의 품질 검증항목과 목표수치는 아래와 같다.

[표2] 품질 지표 및 목표

품질특성	항목명	측정 지표	정량 목표
다양성	클래스 (감굴 부위) 별 라벨 분포	비율	분포확인
	클래스 (키위 부위) 별 라벨 분포	비율	분포확인
	클래스 (병충해) 별 라벨 분포	비율	분포확인
	감굴 초분광 생육 단계 분포	비율	분포확인
	키위 초분광 품종 분포	비율	분포확인
구문정확성	구문 정확도	정확도(%)	99.9% 이상
의미정확성	폴리곤 라벨 정확도	정밀도@IoU0.8	95% 이상
		재현율@IoU0.8	95% 이상
		F1-score	95% 이상
	분류 라벨 정확도	정밀도@IoU0.8	95% 이상
		재현율@IoU0.8	95% 이상
		F1-score	95% 이상
유효성	객체 분류	F1-score	85% 이상

2.2 검증방법

항목별 검증방법은 아래와 같다. 다양성, 구문정확성, 의미정확성 검증은 검증 대상 데이터를 별도의 클라우드 기반 분석 플랫폼에 업로드 한 후 진행된다. 그리고 유효성 검증은 데이터 구축 수행기관이 구성한 클라우드 또는 별도의 로컬환경에서 진행된다.

[표3] 검증항목 별 검증방법 요약

품질특성	검증방법
다양성	<div> <div>검사대상 분석</div> <div>→</div> <div>통계 분석 규칙 마련</div> <div>→</div> <div>검사 결과 리포팅</div> </div> <p>1) 검사 대상 라벨링 데이터 파일의 구조 및 형식을 분석하고 다양성 통계분석에 필요한 속성자 포함 여부 확인 및 보완 조치</p> <p>2) 통계적 다양성 분석에 적용할 분석 규칙 확정</p> <p>3) 자동화 도구(주비투엔 SDQ for AI)를 활용하여 전수 검사</p>
구문 정확성	<div> <div>검사대상 분석</div> <div>→</div> <div>구문 진단 규칙 마련</div> <div>→</div> <div>검사 결과 리포팅</div> </div> <p>1) 검사 대상 라벨링 데이터 파일의 구조 및 형식을 분석하고 구문 정확성 검사 자동화에 적합한 포맷 여부 확인 및 보완 조치</p> <p>2) 라벨링 데이터 구문 정확성 검증에 적용할 구문 진단 규칙 확정</p> <p>3) 자동화 도구(주비투엔 SDQ for AI)를 활용하여 전수 검사</p> <ul style="list-style-type: none"> - (대상 누락) 원본 데이터 식별자 누락 사례 확인 - (파일 오류) 라벨링 데이터 파일 읽기 오류 사례 확인 - (항목 오류) 라벨링 데이터 필수 항목(property) 누락 사례 확인 - (형식 오류) 필수 항목의 값(value) 누락 및 허용 범위 초과 사례 확인
의미 정확성	<p>1) 검사 대상 데이터 샘플링</p> <p>2) 검증항목 특성을 고려하여 아래의 방법 중 하나를 선택하여 평가 - ② 번을 선택</p> <p>① 클라우드 기반 정성평가 ② 라벨링 숙련자 육안검사 ③ 참조 정답값(GT) 제작 ④ 전문가 검사</p> <p>※ 인공지능 학습용 데이터 검사작업을 위해 (주)슈퍼브에이아이 Superb Suite[®]플랫폼 을 활용</p>
유효성	<p>1) 훈련/(검증)/평가용 데이터셋 준비</p> <p>2) 알고리즘을 훈련/(검증) 데이터셋으로 학습하여 인공지능 모델 준비</p> <p>3) 평가용 데이터셋을 인공지능 모델에 입력하여 결과 기록</p> <p>※ 데이터셋, 인공지능 모델 및 구동환경은 수행기업(기관)에서 제공</p>

3. 검증환경

3.1 데이터 검증조건

검증항목 별 품질 검증 대상 데이터의 포맷, 수량 및 측정방법은 아래와 같다.

[표4] 품질검증 대상 데이터 현황 및 품질항목 별 측정방법

품질특성	내용	
다양성	검증대상 데이터	o 데이터 포맷: JSON o 데이터 수량: 병충해 - 40,000건(파일), 초분광 - 100,000건(파일)
	측정 산식	o 각 항목별 통계 분포 확인
구문 정확성	검증대상 데이터	o 데이터 포맷: JSON o 데이터 수량: 병충해 - 40,000건(파일), 1,160,000건(개체) 초분광 - 100,000건(파일), 3,700,000건(개체)
	측정 산식	o 구조 정확도(파일) = (구조 오류를 제외한 파일 수) / (전체 항목 파일 수) o 구조 정확도(항목) = (구조 오류를 제외한 건수) / (전체 항목 건수) - 파일 정확도에서 오류가 없는 파일을 대상으로 전체 항목 계수 o 형식 정확도 = (형식 오류를 제외한 항목 건수) / (전체 항목 건수)
의미 정확성	검증대상 데이터	o 데이터 포맷: jpg(이미지), JSON(라벨) o 데이터 수량: 이미지 2,046장, 라벨 2,046개
	측정 산식	$\text{precision} = \{TP\} / \{TP+FP\}$ $\text{recall} = \{TP\} / \{TP+FN\}$ $F1\text{-score} = 2 * \{\text{precision} * \text{recall}\} / \{\text{precision} + \text{recall}\}$
유효성	검증대상 데이터	o 데이터 포맷: jpg(이미지), JSON(라벨) o 데이터 수량: 이미지 4,002장 라벨 4,002개(객체 분류)
	성능측정 알고리즘	o 학습 모델명: ResNet, DenseNet

3.2 유효성 검증환경

인공지능 유효성 검증 환경 및 학습조건은 아래와 같다. 훈련/(검증)/평가용 데이터셋 및 인공지능 모델은 데이터 구축 수행기관에서 제공한다.

[표5] 유효성 검증 환경

유효성 검증 항목							
항목명	객체 분류						
검증 방법	도커이미지 제출						
목적	과수작물 병충해 분류						
지표	F1 score						
측정 산식	F1 Score = 2*(Recall * Precision) / (Recall + Precision)						
도커 이미지	2_15_HF_dataset_docker_img						
실행 파일명	eval.py						
유효성 검증 환경							
CPU	Intel(R) Core(TM) i9-9900KF						
Memory	128GB						
GPU	NVIDIA TITAN RTX						
Storage	Samsung SSD 970 Evo Plus 1TB						
OS	Windows10						
유효성 검증 모델 학습 및 검증 조건							
개발 언어	Python 3.9.5						
프레임워크	CUDA 11.1 tensorflow 2.5.0						
학습 알고리즘	ResNet, DenseNet						
학습 조건	epoch:100, batch:64, optimizer:Adam						
파일 형식	• 학습 데이터셋: jpg, json • 평가 데이터셋: jpg, json						
전체 구축 데이터 대비 모델에 적용되는 비율	AI모델 사용 이미지 비율(수량) - 과수작물(감귤) 병충해 이미지 촬영: 100% (20,000장 이상) - 과수작물(키위) 병충해 이미지 촬영: 100% (20,000장 이상)						
모델 학습 과정별 데이터 분류 및 비율 정보	• Training / Validation / Test Set 비율 및 수량						
	구분	감귤			키위		
		병충해명	비율	수량	병충해명	비율	수량
	Training Set	정상	80%	4,000	정상	80%	3,998
		궤양병	80%	9,014	점무늬병	80%	6,142
		굴응애	80%	1,452	총채벌레	80%	4,467
		진딧물	80%	1,534	과실무름병	80%	1,389
	Validati on Set	정상	10%	500	정상	10%	501
		궤양병	10%	1,127	점무늬병	10%	768
		굴응애	10%	181	총채벌레	10%	559
		진딧물	10%	192	과실무름병	10%	174
	Test Set	정상	10%	500	정상	10%	501
궤양병		10%	1,127	점무늬병	10%	768	
굴응애		10%	181	총채벌레	10%	559	
진딧물		10%	192	과실무름병	10%	174	
제한사항							

4. 검증결과

항목별 검증 결과는 아래와 같다. 다만 아래의 결과값은 본 결과서에 기술된 검증환경에서 나온 것으로 검증환경이 달라지거나 데이터 구성 변경 등이 적용되는 실제 환경에서는 결과값이 달라질 수 있다.

[표6] 인공지능 학습용 데이터 품질검증 결과 요약

품질특성	항목명	측정 지표	정량 목표		결과값		목표 충족여부
다양성	클래스 (감귤 부위) 별 라벨 분포	비율	과실	분포확인	4,261개	21.31%	분포확인
			잎		15,739개	78.70%	
	클래스 (키위 부위) 별 라벨 분포	비율	과실	분포확인	3,861개	19.31%	
			잎		16,139개	80.70%	
	클래스 (병충해) 별 라벨 분포	비율	골드키위-정상	분포확인	3,386개	8.47%	
			골드키위-점무늬병		1,989개	4.97%	
			골드키위-총채벌레		4,562개	11.41%	
			골드키위-과실무름병		1,737개	4.34%	
			레드키위-정상		1,614개	4.04%	
			레드키위-점무늬병		5,689개	14.22%	
			레드키위-총채벌레		1,023개	2.56%	
			온주밀감-정상		5,000개	12.50%	
			온주밀감-게양병		11,268개	28.17%	
			온주밀감-굴응애		1,814개	4.54%	
			온주밀감-진딧물		1,918개	4.80%	
	감귤 초분광 생육 단계 분포	비율	온주밀감 6	분포확인	13,092개	27.89%	
			온주밀감 7		36,098개	72.19%	
	키위 초분광 품종 분포	비율	골드키위	분포확인	45,706개	91.41%	
			레드키위		4,294개	8.59%	

구문 정확성	구문 정확도	정확도(%)	99.9% 이상	구조(파일)	100%	달성
				구조(항목)	100%	
				형식	100%	
의미 정확성	폴리곤 라벨 정확도	정밀도@ IoU0.8	95% 이상	99.71%	달성	
		재현율@ IoU0.8	95% 이상	100%	달성	
		F1-score	95% 이상	99.85%	달성	
	분류 라벨 정확도	정밀도@ IoU0.8	95% 이상	99.65%	달성	
		재현율@ IoU0.8	95% 이상	100%	달성	
		F1-score	95% 이상	99.82%	달성	
유효성	객체 분류	F1-score	85% 이상	98.45%	달성	

※ 상세결과는 별도로 제공하는 자료(엑셀시트)에서 확인 가능

5. 기타사항

하기와 같은 구문 오류에 대해 보완 조치 되었음을 확인하였다.

스키마에 허용되지 않는 프로퍼티 존재
<pre>"Annotations": { "ANTN_ID": "1", "ANTN_TY": "polygon", "OBJECT_CLASS_CODE": "감귤_굴응애", "BNDBX": "[33113311461158516931828191811071111951124411296113401137511380113961141011407113341129111196110831892117616711616159915521425],[41514471497155316041624161616051590156715251503149514991486148414651396136313341350135513671372138513871405]" }</pre>
보완 예시
<pre>"Annotations": { "ANTN_ID": "1", "ANTN_TY": "polygon", "OBJECT_CLASS_CODE": "감귤_굴응애", "ANTN_PT": "[33113311461158516931828191811071111951124411296113401137511380113961141011407113341129111196110831892117616711616159915521425],[41514471497155316041624161616051590156715251503149514991486148414651396136313341350135513671372138513871405]" }</pre>

항목명 내 문자열(3..37) 불일치
<pre>"Hyperspectral": { "HS_NON_BRX": null, "HS_NON_DM": null, "HS_BRX": "13.8", "HS_DM": "21.36", "HS_CHROMA": "101.15", "HS_HN": "3..37", "HS_ACIDITY": "1.27" }</pre>
보완 예시
<pre>"Hyperspectral": { "HS_NON_BRX": null, "HS_NON_DM": null, "HS_BRX": "13.8", "HS_DM": "21.36", "HS_CHROMA": "101.15", "HS_HN": "3.37", "HS_ACIDITY": "1.27" }</pre>